Journal of Trends in Computer Science and Smart Technology (ISSN: 2582-4104)
www.irojournals.com/tcsst/

Cross-Lingual Attention-based Mechanism for Speech Emotion Recognition

Tummala Vamsi Aditya[1], Swarna Kuchibhotla[2], Devi Venkata Revathi Poduri[3], Hima Deepthi Vankayalapati[4]

[1-3]Department of Computer Science and Engineering, Koneru Lakshmaiah Education Foundation, Vaddeswaram, India.
[4]Department of Artificial Intelligence, Mukesh Patel School of Technology Management and Engineering, NMIMS University, Mumbai, India.

Email: [1]vamsiaditya9835@gmail.com, [2]drkswarna@kluniversity.in, [3]pdvrevathi@gmail.com, [4]nanideepthi@gmail.com

Abstract

Speech emotion recognition is one of the most emerging areas for emotion detection that may fall within the scope of affective computing. In this particular case, emotional speech files of spoken words delivered during verbal communication are of interest. The emotions of speech are investigated through sound and emotion in speech and are modeled through machine learning. Through machine learning, we performed a series of experiments on datasets like RAVDESS, TESS, SAVEE, and EMO-DB, which lean toward the objective that a Recurrent Neural Network (RNN) and (CLAF-SER): The Cross-Lingual Attention-Based Adversarial Framework for SER would be able to detect and classify such emotions as sadness, anger, happiness, neutrality, and fear. Features such as MFCC, LPCC, pitch, energy, and chroma were extracted before implementing the RNN. Through this model, TESS achieved the highest accuracy among the other datasets. However, CLAF-SER gives the best performance when all datasets are combined.

Keywords: Speech Emotion Recognition (SER), RNN (Recurrent Neural Network), CLAF-SER (Cross-Lingual Attention-based Adversarial Framework for SER), SAVEE (Surrey Audio-Visual Expressed Emotion Database), RAVDESS (Ryerson Audio-Visual Database of Emotional Speech and Song), TESS (Toronto Emotional Speech Set), EMO-DB (Berlin Database of Emotional Speech), MFCC (Mel-Frequency Cepstral Coefficients), LPCC (Linear Prediction-based Cepstral Coefficients), Pitch, Energy, Chroma.

1. Introduction

A world in which computers can not only interpret what people say but also comprehend the emotions that accompany every single word. This beautiful image is the essence of SER, which is the cutting-edge field of AI that tackles emotional speech from an entirely different perspective [21]. During this phase, SER devices are capable of identifying more than the usual assortment of words and considering other complex aspects of the voice like pitch changes, energy fluctuations, and micro-intonations [15]. Additionally, it can completely change the way to perceive technology, bringing us closer to it and ultimately improving our existence in other unrelated areas [18].

DOI: https://doi.org/10.36548/jtcsst.2025.3.003
Received: 20.06.2025, received in revised form: 21.07.2025, accepted: 05.08.2025, published: 16.08.2025

The main deep learning principle in a model used for SER is applied, which is the skill of feature extraction. In here, the speech takes into account vital attributes that convey emotional information. Such components work as an "emotional fingerprint" analogue because they create a unique sound for each sentiment. Consequently, in the spectrum domain,

their features have been characterized as roughness, noise, timbre, and entropy, respectively [25]. They serve as a basis for communication between man and computer (namely, in virtual space, or the internet). On the downside, the tough task remains in the process of unveiling the complicated web of connections that links mental health to its physical manifestation [20].

Distinct from classic models, RNNs excel in handling sequential data, such as speech. Imagine a sentence as a chain of words. RNNs can thus determine the structure of the sentence and check the exact meaning of each word, either due to the word itself or due to the words around it [23]. The possibilities of SER technology start from scientific discoveries and extend to practical applications with numerous solutions and civic importance [16].

Revolutionizing customer service and enhancing education can be achieved through the implementation of SERS (Sentiment Emotion Recognition Systems). In customer service, picture a call center equipped with a SERS system that can identify a customer's distress in real-time, allowing agents to address their concerns with empathy and intention, ultimately diffusing tense situations [24]. By adopting a personalized approach, companies can create memorable customer experiences, resulting in happier, more loyal customers. Similarly, in education, SERS can be integrated into educational tools to measure student engagement and emotional responses, enabling adjustments to teaching methods and curricula to meet individual needs. This tailored approach fosters a more effective and productive learning environment, enhancing interpersonal interactions and promoting better educational outcomes [26].

Optimizing healthcare and elevating human-computer interaction can be significantly advanced through the implementation of Sentiment Emotion Recognition (SER) systems. In healthcare, these systems can recognize and analyze communication between patients and healthcare providers, allowing for a deeper understanding of mood changes [17]. This information can form the basis for guidelines and protocols aimed at delivering more humanized care, enabling earlier detection of mental health concerns and timely interventions. Similarly, in the realm of human-computer interaction, virtual assistants and robots leverage SER technology to respond to users and adapt to their emotional states [22]. This interaction creates a more genuine and engaging experience, making the way for a future where machines not only understand our instructions but also interpret our emotions, enhancing the overall user experience [19].

2. Related Work

Singh et.al. proposed a deep learning model, which is attention-based by combining a 2D Convolutional Neural Network (CNN) and a long short-term memory (LSTM) network. They used this model to find the best features that sclassify. emotions precisely for their custom dataset, which is a combination of SAVEE, RAVDESS, and TESS. The overall accuracy achieved is 90% [1]. Sun C et.al. proposed a SER framework, which they named IMEMD-CRNN, based on combining a convolutional recurrent neural network (CRNN) and an upgraded version of the masking signal-based EMD (IMEMD). Even though there are various

EMD methods, they have some problems, like long computation time and residual noise. To overcome these issues, they proposed IMEMD-CRNN that involved only the TESS and EMO-DB databases, achieving accuracies of 100% and 93% [2].

Alluhaidan Ala Saleh et.al. they proposed a CNN to build the SER model and hybrid features i.e; combination of MFCC and time domain features were given to this model. The algorithms used in this paper are CNN, KNN (k nearest neighbor), RF (Random Forest), NB (Naive Bayes), and SVM (Support Vector Machine). Through experiments, they concluded that CNN was the best among others, obtaining a 97% accuracy on the EMO-DB dataset, achieving 92.6% on SAVEE and 91.4% accuracy on RAVDESS using their CNN model [3].

Saumard Matthieu et.al. used the SVM algorithm for this paper with MFCC as multivariate functional data. For the EMO-DB database, they achieved 85.8%, and for the IEMOCAP database, they achieved an accuracy of 65.2%. They mentioned that their method of approach probably reduces the learning time and makes it more efficient and practical for real-world problems [4].

Md. Rayhan Ahmed et.al. used a model with a combination of 3 different architectures, which are 1DCNN, LSTM(Long Short Term Memory), GRU(Gated Recurrent Unit), focusing on extracting local as well as global representation of speech signals from TESS, RAVDESS, SAVEE, EMO-DB, CREMA-D(Crowd-sourced Emotional Multi-modal Actors Dataset) datasets with accuracies of 99.46%, 93.22%, 95.62%, 95.42%, 90.47% [5]. Bagadi Kesava Rao et.al. their aimed to investigate how the feature selection meta-heuristic approaches impact emotion recognition in speech. They used the RAVDESS dataset, employing Equilibrium Optimization (EO) and Cuckoo Search (CS) for feature selection, along with an SVM classifier to achieve accuracies of 89.64% and 92.71% accuracies [6].

Taiba Majid Wani et al. reviewed several studies on Speech Emotion Recognition (SER) that used datasets like EMO-DB, SAVEE, SEMAINE, RECOLA, and IEMOCAP. They noted that different models, such as SVM, HMM, Naive Bayes, GMM, LSTM, and RBM, were tested on these databases, with their accuracies compared across studies [7]. In another work, Langari Shadi et al. introduced a feature extraction method based on adaptive time-frequency coefficients using the SAVEE, EMO-DB, and PDREC datasets. By combining SVM with a GA-CS feature selection algorithm, they achieved 80% accuracy on SAVEE, 91.46% on PDREC, and 97.57% on EMO-DB [8].

Ruhul Amin Khalil et.al., their work is a review of different papers worked on SER and mentioned that commonly deep learning techniques such as RNN, DBN (Deep Belief Networks), CNN, and autoencoders are applied to datasets like SAVEE, CREMA-DB, RAVDESS, and many others. However, the accuracies were not disclosed in this paper; instead, they stated that these methods make model training easier as well as improve the efficiency of shared weights [9]. Akçay Mehmet Berkehan et.al. provided a detailed survey of current literature based on different papers involved in SER. They made a study on a wide range of datasets and models used on those datasets [10].

While speech emotion recognition (SER) has come a long way, there are still several key challenges, especially when it comes to handling real-world, multilingual scenarios. Most SER models are trained using data from just one language, which means they often struggle to perform well when exposed to speech in other languages. They tend to rely on language-specific sound patterns, limiting their flexibility. Additionally, many existing approaches don't fully address the issue of domain adaptation, so their performance drops when

tested on new datasets or languages. Another gap lies in the way features are used; prior models often depend on broad acoustic features like MFCCs but miss out on important phonetic-level details that are shared across languages. Attention mechanisms have been used in some models, but usually not in a way that's deeply integrated with both temporal modeling (like BiLSTMs) and language adaptation. Finally, many powerful models are too large or complex to run on mobile or edge devices in real time. To address all of these issues, we propose CLAF-SER, a unified framework that brings together several techniques. It uses BiLSTM with multi-head attention to better capture emotional context over time, applies a Gradient Reversal Layer (GRL) to help the model ignore language differences during training, includes a Conv1D block to extract phonetic patterns, and adds language embeddings to make the system more adaptive. On top of that, it's designed to be lightweight and efficient, making it suitable for real-time use on edge devices. With this combination, CLAF-SER addresses major limitations in current SER research and provides a more robust, flexible, and deployable solution.

3. Proposed Work

Since this paper uses both RNN and CLAF-SER, it's important to first understand why RNNs are used and their architecture before exploring CLAF-SER. A Recurrent Neural Network (RNN) is a type of neural network designed to work with sequential data, making it suitable for tasks like time series analysis, language modeling, and speech recognition. Unlike Feed-Forward Neural Networks (FNNs), which process each input independently, RNNs have looped connections that allow them to retain information from previous inputs. These feedback loops help the network remember past patterns and use that context for future predictions, which is why they are called 'recurrent.' This ability makes RNNs powerful for tasks where order and context matter, such as translation, speech recognition, and other time-dependent applications. In contrast, FNNs cannot handle temporal relationships in data, giving RNNs a significant advantage for sequential processing.

3.1 Features Extracted

3.1.1 MFCC

MFCC (Mel Frequency Cepstral Coefficients) defines a new audio feature extraction method using the auditory characteristics of humans. The audio signals, although using manipulations that are based on the Fourier Transform, Mel scale filter bank, and Discrete Cosine Transform, represent audio signals compactly and lend themselves to a range of applications, including speech recognition and speaker identification.

$$C_k = \sum_{m=0}^{M-1} \log(S_m) \cdot \cos\left[(\pi k (2m+1)) / (2M) \right] \quad (1)$$

where Ck are the MFCCs, M is the number of Mel filter bank coefficients, and k is the index of the MFCC, S is the Mel-scaled power spectrum, m is the corresponding Mel frequency.

3.1.2 LPCC

This is a technique based on linear predictive coding termed LPCC, whose quality is highly regarded in the extraction of features, the spectral envelope of a signal being represented as a spectral envelope. Among the various applications are speech processing, including

recognition and synthesis, where this feature is generally used because of its compact and efficient nature.

$$c[m] = a[m] + \sum_{(k=1)}^{(m-1)} (k/m) \cdot c[k] \cdot a[m-k], \text{ if } m \leq p \quad (2)$$

$$c[m] = \sum_{(k=1)}^{(p)} (k/m) \cdot c[k] \cdot a[m-k], \text{ if } m > p \quad (3)$$

where c[m] are the LPCCs, a[m] are the LPC coefficients.

3.1.3 Pitch

Pitch is the most basic frequency characteristic and refers to the vibration rate of the vocal folds. It is vital for conveying emotion in speech. It plays a significant role in Speech Emotion Recognition, especially since pitch variations correspond to different emotional states, such as anger, sadness, or happiness.

$$R(\tau) = \sum_{(n=0)}^{(N-1)} x[n] \cdot x[n+\tau] \quad (4)$$

$$F_0 = 1/T_0, \quad T_0 = \arg\max_{(\tau_{min} \leq \tau \leq \tau_{max})} R(\tau) \quad (5)$$

where x[n] is the speech signal, τ is the lag, N is the frame length, F0 is the fundamental frequency, T0 is the quefrency corresponding to the highest peak in the cepstrum.

3.1.4 Energy

Energy is what, in reality, signifies the intensity of speech production and is reflected in loudness or amplitude fluctuations. In Speech Emotion Recognition, an integral part is that different emotions, like anger or excitement, usually generate a lot of energy, while sadness or boredom are generally low energy.

$$E = \sum_{(n=1)}^{(N)} |x[n]|^2, \quad \text{or} \quad \text{LogEnergy} = \log\left(\sum_{(n=1)}^{(N)} |x[n]|^2 + \epsilon \right) \quad (6)$$

where E is the Energy of the frame, x[n] is Speech signal sample at index n, N is the Number of samples in the frame, ε is a small constant to avoid numerical issues.

3.1.5 Chroma

Chroma features are used to represent how energy is distributed over the various pitch classes in music, focusing on the harmonic and tonal characteristics of sound. In SER, they can also be used to analyze the timbre of the voice and understand its mood to differentiate emotions using variations in tonality. Compute the magnitude spectrum:

$$|X(f)| = \text{FFT}(x[n]) \quad (7)$$

Identify the frequency bins from the FFT that correspond to each pitch class. Map frequencies to chroma bins:

$$C[k] = \sum_{(f \in \text{Bin}(k))} |X(f)| \quad (8)$$

where C[k]: Chroma energy for pitch class k.

Normalize the chroma vector:

$$\hat{C}[k] = C[k] / \sum_{(j=1)}^{(12)} C[j] \quad (9)$$

where Cˆ[k] is the normalized chroma feature.

3.2 Datasets

In this paper, four datasets are used, which are famous for the classification of emotions.

3.2.1 TESS Data

It is a dataset developed to investigate emotional expression in speech. It contains recordings of two females of different ages, one younger and one older, who read words with different emotional tones such as disgust, fear, anger, happiness, surprise, neutrality, and sadness. Each speaker speaks 200 words in a neutral tone of voice with such phrases as "Say the word." This allows for the analysis of expressive intentions of different age groups as well as emotional intention. TESS is valuable and is primarily employed for emotion recognition tasks, and it can be used in developing models that recognize emotions in spoken language [14].

Table 1. Different datasets used in SER

Name of Dataset	Number of Files	File Format	Emotions
SAVEE	480	WAV format	happiness(60), neutral(120), fear(60), sadness(60), surprise(60), anger(60) and disgust(60)
TESS	2800	WAV format	disgust(400), happiness(400), anger(400), surprise(400), sadness(400), neutral(400) and fear(400)
RAVDESS	1440	WAV format	disgust(192), calm(192), fear(192),neutral(96), sad(192), happy(192), surprise(192) and anger(192)
EMO-DB	535	WAV format	happiness(71), disgust(46), fear(69), anger(127), boredom(81), neutral(79) and sadness(62)

3.2.2 RAVDESS Data

It is a database intended for the study of emotion detection in spoken words and songs. Recordings were made from a total of 24 actors who received training (12 male and 12 female), each portraying a range of eight emotions: sad, angry, calm, happy, disgust, fearful, surprise, and neutral. The emotions are produced at varying degrees of intensity and feature both speech and sung utterances; this offers an extensive range of varied emotional expressions. RAVDESS is commonly used for training and testing models that recognize emotions in audio and visual data, making it valuable for studies in affective computing and human-computer interaction [11].

3.2.3 SAVEE Data

It is a tool developed specifically for emotion recognition in speech research. The voice database was developed at the University of Surrey and consists of recordings of 4 different male actors producing 7 emotions: sadness, surprise, fear, happiness, anger, disgust, and neutral. The database consists of 480 British English sentences and is designed in such a way as to capture all emotional variances. SAVEE also contains visual data in addition to audio, which helps in building emotion recognition systems that diagnose in a single or two modes. It's often used in affective computing research to improve human-computer interaction by enabling systems to better understand human emotions [13].

3.2.4 EMO-DB Data

It is an emotion research-oriented dataset in German. It contains recordings of ten native German speakers (5 female, 5 male) with emotions such as fear, happiness, anger, boredom, disgust, neutrality, and sadness as shown in Table 1. Every file recorded in this collection was listened to by experts in order to ascertain the intended emotion, ensuring the validity of the data collection's EMO-DB has many high-quality audio examples and is very popular for building and testing emotion recognition systems, particularly within Spanish-speaking countries, and therefore is useful in affective computing and speech-based emotion detection [12].

3.3 RNN Architecture

These are types of neural networks that have hidden layers and allow the inclusion of previous outputs as current inputs. The creation of an RNN generally consists of the following process:

1. Input Layer: This is the main layer that receives the first data element of the sequence, for example, a complete sentence ready to input the first word as a vector.

2. Hidden Layer: The primary part of an RNN is this hidden layer, which consists of many interconnected neurons. A neuron receives both the current input and the information coming from the previous layer's hidden state. This information, coming from the previous hidden layer state, is termed 'state', which is the only element that remembers what has occurred with earlier inputs and helps to adjust the present input to its relevant context. Thus, using formula (1) updates the hidden state.

3. Activation Function: This ensures that there will be some non-linearity in the net; otherwise, it will not be able to grasp more complicated relations. It is applied to the integrated input at the current input layer and the states of earlier hidden layers before passing the integrated input to the next process.

4. Output Layer: The layer that performs the specified function and results in the output of the network. For example, the output will be the next word in a sequence of words provided as in the case of a language model. Using formula (2) output calculation is performed.

5. Recurrent Connection: This is another distinguishing characteristic of RNNs, whereby connections can also be made within a hidden layer. Along with this connection, the network retains the previous transient pace horizontal state

information and makes it available for the next transient pace step. It is similar to relaying, where the earlier runners have to pass something to the next one.

6. Equations: Hidden state update:

$$h_t = \sigma(W_h \cdot h_{t-1} + W_x \cdot x_t + b_h) \tag{10}$$

Output Calculation:

$$y_t = W_y \cdot h_t + b_y \tag{11}$$

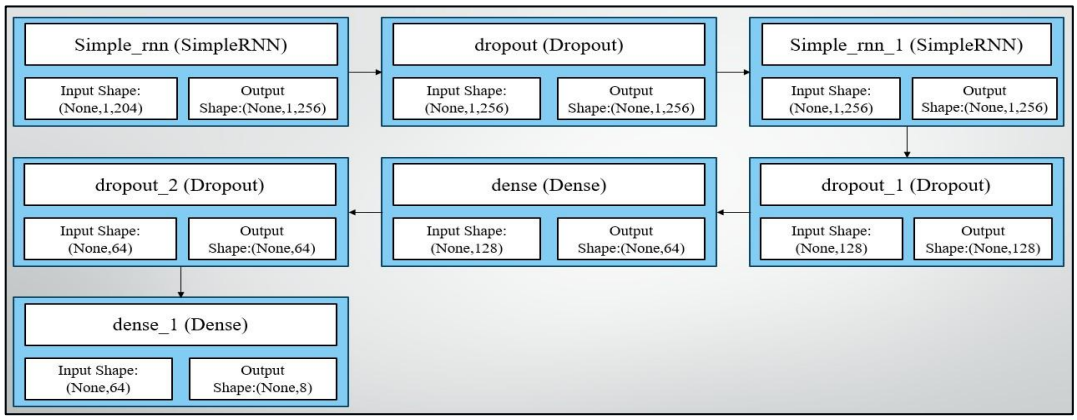

Figure 1. RNN Architecture Followed to Implement SER

Where σ is the activation function (e.g., tanh or ReLU), W_x and W_h are weight matrices, and b_h and b_y are biases.

The architecture of a sequential neural network, i.e, RNN, is depicted in the Figure 1. It begins with a SimpleRNN layer, which contains 256 units, learns inputs that have the shape of (None, 1, 204), and continues with a Dropout layer to avoid overfitting. Another 'SimpleRNN' layer with 128 units is included afterward, which is again succeeded by another dropout layer. The Dense layer with 64 units performs dimensionality reduction, after which a final dropout layer and a dense layer with 8 output units, probably corresponding to the classification tasks, are applied. In order to track the data transfer throughout the network, images illustrate the shape of every layer's inputs and outputs.

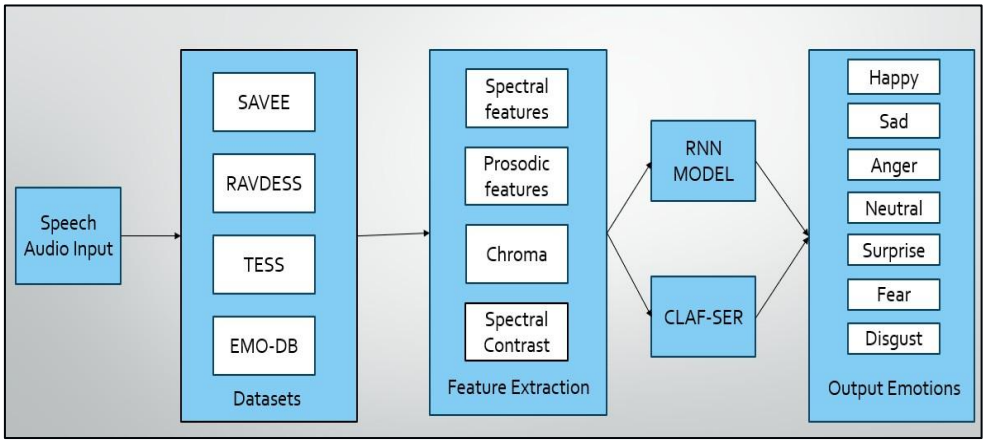

Figure 2. Block Diagram of Proposed Work

From Figure 2, we can see the flow followed for this paper. The inputs are taken as speech audio, where selecting the datasets with different audio files is crucial for this project. In that context, the SAVEE, RAVDESS, TESS, and EMO-DB datasets were selected for this project. These datasets consist of audio files with different types of emotions, such as anger, sadness, happiness, neutrality, disgust, calmness, surprise, and fear. These files are then given as input. From the given input files, feature extraction is performed, and the extracted features are spectral features, which consist of MFCC, LPCC, and prosodic features like pitch, energy, and chroma. These extracted features are stored as data, which is then split into training and testing datasets. To this data, RNN and CLAF-SER were applied in order to obtain the output.

3.4 Cross-Lingual Attention-based Adversarial Framework for Speech Emotion Recognition (CLAF-SER)

The Cross-Lingual Attention-based Adversarial Framework for Speech Emotion Recognition (CLAF-SER) aims to capture emotional patterns that work across different languages. It combines multi-head attention, adversarial domain adaptation, and phonetic feature extraction for better generalization. The model processes features like MFCCs, pitch, chroma, energy, and spectral contrast using a BiLSTM to learn sequential patterns, while multi-head attention highlights key emotional moments. A gradient reversal layer (GRL) removes language-specific traits, and a parallel 1D CNN extracts language-independent phonetic cues. These features are fused and passed through a classifier with normalization and dropout for stable predictions. By jointly training on emotion classification and domain confusion, the framework ensures accurate and language-robust emotion recognition. The architecture consists of key components as follows:

3.4.1 Audio Preprocessing and Feature Extraction

The audio preprocessing and feature extraction stage begins by resampling and normalizing each audio file to ensure consistency across different datasets. Using Librosa-based tools, several key features are extracted from the waveform, including MFCCs to capture short-term spectral characteristics, pitch to represent prosodic variations, RMS energy for measuring intensity, chroma for harmonic content, and spectral contrast to reflect timbral properties. These diverse features are then concatenated and padded to a fixed sequence length, resulting in a uniform input tensor. This tensor is structured to align with the expected input format of the LSTM layer, serving as the starting point for the CLAF-SER model's processing pipeline. Each audio utterance is transformed into a fixed-length time-series feature sequence

$$X = [x1, x2, \ldots, xT\ ] \in RT \times d \tag{12}$$

where T is the number of time steps (padded/truncated), and d = 34 is the feature dimension.

3.4.2 BiLSTM Layer (Temporal Modeling)

After preprocessing and padding, the audio features are fed into a Bidirectional LSTM (BiLSTM) layer. Unlike standard LSTMs that process data in one direction, BiLSTMs analyze sequences both forward and backward, capturing emotional patterns from the entire utterance. This is crucial for multilingual emotion recognition, where cues may appear at the end of a sentence or depend on future context, such as sarcasm, pauses, or hesitation. By using past and future information at each step, the BiLSTM adapts to language-specific variations in prosody

and syntax. Its output feeds into the multi-head attention module as the query, key, and value, while also serving as input for phonetic feature extraction through the convolutional branch. This dual representation strengthens the model's ability to detect emotionally rich moments across different languages and speaking styles, even in mixed-language scenarios. A bidirectional LSTM encodes the sequence in both directions.

$$H = \text{BiLSTM}(X) \in RT \times 2h \tag{13}$$

where h = 128 is the hidden size per direction.

3.4.3 Multi-Head Attention Mechanism

After the BiLSTM processes the audio sequence, its output is passed through a multi-head attention mechanism. This layer plays a crucial role in helping the model focus on the most emotionally meaningful parts of the utterance. Rather than treating every time step equally, multi-head attention dynamically weighs different moments in the sequence, such as stressed syllables, sharp pitch changes, or emotionally charged pauses, based on their relevance to the overall emotion. By doing so, the model becomes more sensitive to subtle emotional cues that may not be captured by fixed window analysis alone. The outputs from all attention heads are then combined and averaged using mean pooling, resulting in a single, detailed vector. This global representation captures the emotional essence of the entire utterance and serves as a critical input for both the final emotion classification and the domain adaptation branch of the CLAF-SER model.

Multi-head attention is applied over the temporal outputs:

$$Z = \text{MultiHeadAttention}(H, H, H) \in RT \times 2h \tag{14}$$

$$\mathbf{f}_{\text{attn}} = (1 / T) \sum_{(t=1)}^{(T)} \mathbf{z}_t \in \mathbb{R}^{2h} \tag{15}$$

Along with attention-based processing, CLAF-SER uses a dedicated phonetic feature extraction branch to capture subtle, low-level speech patterns shared across languages. The input features are reshaped for convolutional processing and passed through a 1D convolutional layer followed by adaptive average pooling. This helps the model learn compact phonetic cues like frequency transitions and articulation patterns that LSTMs may miss. These features are more stable across languages, making them crucial for multilingual emotion recognition. The resulting 16-dimensional vector is then concatenated with the attention module's output, combining global emotional context with fine-grained acoustic details for more accurate predictions.

In parallel, phonetic cues are extracted using 1D convolution:

$$Xconv = \text{Conv1D}(XT) \in RC \times 1 \tag{16}$$

$$fphon = \text{Flatten}(Xconv) \in RC \tag{17}$$

3.4.5 Language Embedding Adaptation

To help the model adapt effectively to different languages, CLAF-SER incorporates a language embedding mechanism. During training, each audio sample is labeled with a language or dataset ID, such as RAVDESS, EMO-DB, SAVEE, or TESS, and this ID is converted into a learnable embedding vector. These embeddings serve as a compact summary

of language-specific characteristics, such as prosody, speaking style, and recording conditions. By adding the language embedding to the output of the attention module, the model subtly adjusts its internal representation depending on the language of the input. This helps it account for cross-lingual differences and supports transfer learning across datasets. In essence, the language embedding acts as a form of guided adaptation, giving the model contextual awareness of the speech's linguistic background without requiring manual tuning for each language.

Each input sample has a language ID l ∈ {0, 1, 2, 3}, mapped to a trainable embedding:

$$el = Embed(l) \in R2h \tag{18}$$

$$flang\text{-}adapted = fattn + el \tag{19}$$

3.4.6 Gradient Reversal Layer (Domain Adversarial Training)

To make CLAF-SER effective across multiple languages, it uses an adversarial learning component with a Gradient Reversal Layer (GRL). After combining attention-based emotional cues with language embeddings, the shared representation passes through the GRL, which flips gradient directions during backpropagation. This penalizes the model if it becomes too good at identifying the input language, pushing it to learn language-agnostic emotional features instead. A domain classifier is trained adversarially to ensure the model focuses only on emotions while ignoring language-specific traits. This approach helps CLAF-SER generalize across different datasets and languages, making it highly robust for multilingual and real-world emotion recognition tasks.

A Gradient Reversal Layer (GRL) promotes language-invariant features:

$$\tilde{f} = GRL(flang\text{-}adapted) \tag{20}$$

$$\hat{d} = Softmax(MLPdomain(\tilde{f})) \in RL \tag{21}$$

where L = 4 language domains.

3.4.7 Emotion Classification

In the final stage of CLAF-SER, the model merges the attention-weighted emotional context with phonetic features into one unified vector. This is passed through a dense classifier with fully connected layers, layer normalization, ReLU activation, and dropout to stabilize training and prevent overfitting. The classifier outputs logits representing the probabilities of eight emotion classes, such as anger, happiness, sadness, and surprise. These predictions are used to compute the emotion classification loss during training. By combining multiple information streams into a single decision layer, CLAF-SER achieves accurate and robust emotion recognition across different languages.

Final emotion-specific features are obtained by concatenating adapted and phonetic vectors:

$$ffinal = [flang\text{-}adapted; fphon] \in R2h+C \tag{22}$$

$$\hat{y} = Softmax(MLPemotion(ffinal)) \in RK \tag{23}$$

where K = 8 emotion classes.

3.4.8 Loss Function

CLAF-SER is trained using a joint loss that balances emotion recognition and domain adaptation. For emotion classification, a standard cross-entropy loss is applied based on true emotion labels. In parallel, a domain loss also computed using cross-entropy is used to distinguish between language domains, but it is optimized via a Gradient Reversal Layer (GRL) to encourage the model to learn language-invariant features. To ensure a smooth trade-off, a dynamic weighting factor α is introduced, which increases over training epochs. This allows the model to focus on learning emotion features early on and gradually emphasize domain generalization, leading to more robust cross-lingual performance.

The joint loss function is defined as:

$$L_{total} = L_{emotion} + \alpha \cdot L_{domain} \quad (24)$$

$$L_{emotion} = CrossEntropy(\hat{y}, y) \quad (25)$$

$$L_{domain} = CrossEntropy(\hat{d}, d) \quad (26)$$

Here, $\alpha \in [0, 1]$ is a dynamic weight linearly increased over training epochs.

4. Results and Discussion

The novel aspect of CLAF-SER is that it is the first speech emotion recognition model to integrate BiLSTM-based temporal modeling, multi-head attention, phonetic feature convolution, and adversarial language adaptation with language embeddings in a single unified architecture for robust cross-lingual emotion recognition. Unlike prior work that tackles speaker or dataset mismatch, CLAF-SER directly addresses language variability, enabling it to generalize even across unseen languages or code-switched speech.

CLAF-SER demonstrates partial adaptability to code-switching and intra-sentence language mixing scenarios. While the current model architecture assigns a single language ID per utterance, its adversarial domain adaptation strategy via a Gradient Reversal Layer enables the extraction of language-invariant emotional representations. Additionally, the use of multi-head attention across temporal frames allows the model to focus on emotionally salient segments, regardless of linguistic content.

4.1 Test and Train results of SAVEE

The training and testing loss curves of an RNN model built on the SAVEE dataset are plotted in Figure 3 above with epochs on the x axis and loss on the y axis. At first, both losses are pretty high, showing that the model fails to learn any discriminative features. Over the first few epochs since the start of training, the training loss (shown with a blue line) drops abruptly, as does the testing loss (shown with an orange line), which is a good sign for generalization. After close to 20 epochs, both losses level off with only slight variances in their values, and the corresponding curves stay near each other, indicating that there is a good fit and neither overfitting nor underfitting occurs. This behavior of the RNN model suggests that the model can extract the features necessary for emotion classification very well.

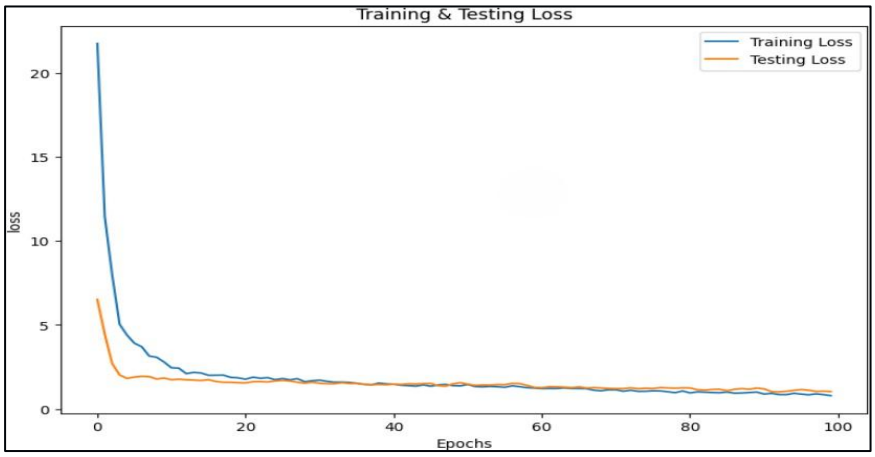

Figure 3. Training and Testing Loss for SAVEE

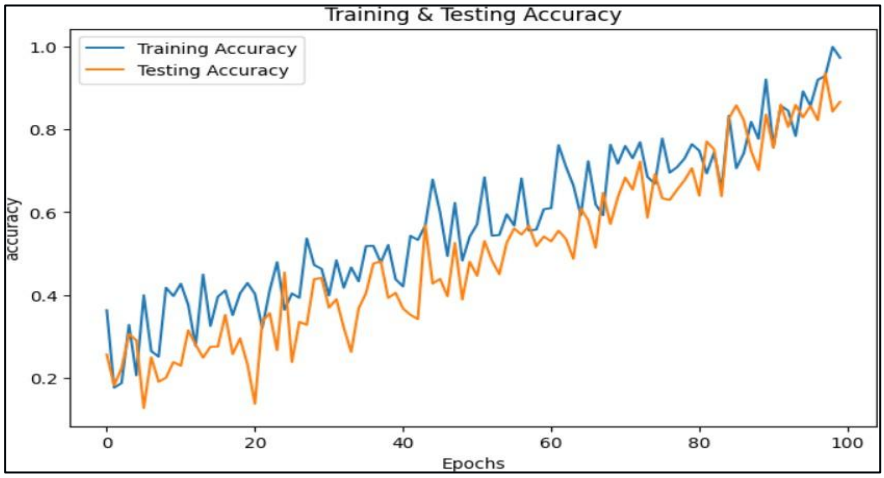

Figure 4. Training and Testing Accuracy for SAVEE

Figure 4 depicts the training and testing performance of an RNN network, which was trained on the SAVEE dataset for 100 epochs, with epochs represented on the x-axis and accuracy represented on the y-axis. It is observed that both accuracies start at a low rate, indicating that the model is not able to accurately recognize the patterns, and classification improves with time. The testing accuracy (orange line) is very unstable in the earlier epochs and becomes more stable in the later epochs, resulting in an initial dip in performance generalization. Training accuracy at the end of training is estimated to be about 97% while testing accuracy is approximately 86 %. This shows that even though the model can extract important characteristics of the training data, it struggles with generalization to unseen data.

Table 2. Performance Metrics for Different Emotions (SAVEE)

Emotion	Precision	Recall	F1-score
Angry	0.89	0.87	0.88
Disgust	0.95	0.75	0.78
Fearful	0.80	0.76	0.74
Happy	0.85	0.84	0.86
Neutral	0.90	1.00	0.95

Sad	0.79	0.87	0.86
Surprised	0.78	0.79	0.77
Accuracy			0.86
Macro Avg	0.85	0.84	0.83
Weighted Avg	0.86	0.85	0.84

The classification metrics regarding a Recurrent Neural Network model deployed on the SAVEE dataset that considers emotions like Angry, Disgust, Fearful, Happy, Neutral, Sad, and Surprised are depicted in Table 2. The model is able to achieve a perfect recall for Neutral, where the degree of separation between the target and non-target classes is the highest. The F1-score is 0.95. The model provides an overall accuracy of 0.86, coupled with a macro average F1 score of 0.83, indicating average performance considering the variability across emotions. The weighted averages indicate marginally improved performance, meaning that the model tends to perform better on the more common classes, like Neutral.

4.2 Test and Train Results of RAVDESS

The training and testing loss curves of the RNN trained on the RAVDESS emotional speech dataset are shown in Figure 5. Both losses start high and gradually decrease, indicating that the model is learning to recognize emotions effectively. Around the 20th epoch, the testing loss stabilizes near 0.2 with slight fluctuations, suggesting good generalization and minimal overfitting. The small gap between training and testing losses reflects solid model training, though minor tuning could further improve performance. Overall, the results show that the RNN effectively captures the sequential patterns in the RAVDESS data for emotion recognition.

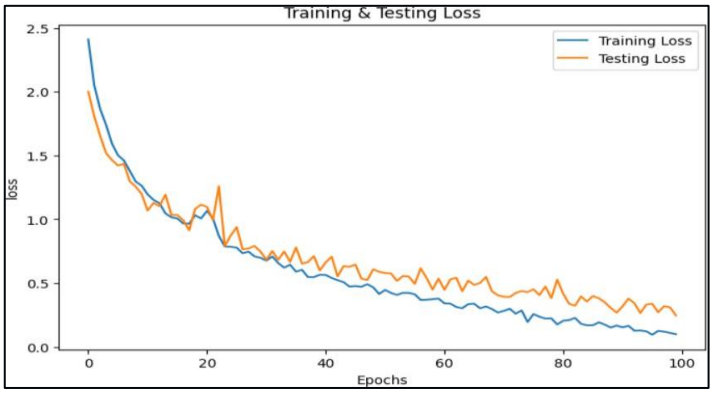

Figure 5. Training and Testing Loss for RAVDESS

Training and testing accuracy curves of the Recurrent Neural Network model trained on the RAVDESS dataset are presented in Figure 6. The accuracies obtained initially are low but progressively increase throughout the training stage of 100 epochs; however, training accuracy went above 90% while testing accuracy stabilized at around 80-85%. The gap between the training and testing accuracy indicates that there is not much overfitting occurring; however, testing accuracy was more variable due to unseen emotional data. This analysis points out that the RNN model is moderately effective in identifying emotional content in the voices in the RAVDESS database.

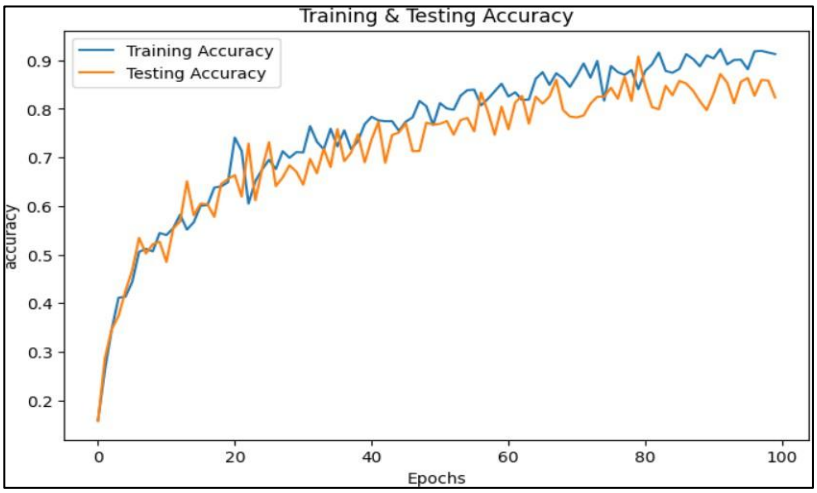

Figure 6. Training and Testing Accuracy for RAVDESS

In Table 3, classification metrics are represented for an RNN model that was built and tested on the RAVDESS database for emotion classification using metrics such as F1-scores, Recall, and Precision. The model performs rather well on emotions like 'Angry' and 'Surprised' with respective F1-scores of 0.89, but performs poorly on 'Neutral' and 'Sad' (F1-scores of 0.77 and 0.76, respectively), probably due to very slight acoustic differences. The Macro and Weighted Averages give insight into the overall performance, and an accuracy of around 82% suggests reasonable performance.

Table 3. Performance Metrics for Different Emotions (RAVDESS)

Emotion	Precision	Recall	F1-score
Angry	0.89	0.84	0.89
Calm	0.87	0.82	0.87
Disgust	0.74	0.76	0.81
Fearful	0.72	0.81	0.79
Happy	0.83	0.79	0.78
Neutral	0.77	0.75	0.77
Sad	0.81	0.77	0.76
Surprised	0.85	0.87	0.89
Accuracy			0.82
Macro Avg	0.81	0.80	0.80
Weighted Avg	0.82	0.82	0.81

4.3 Test and Train Results of TESS

In this case, the model is an RNN trained on the TESS dataset, with its training and testing loss curves shown in Figure 7. The x-axis represents epochs, while the y-axis shows loss values, indicating the model's error. Initially, both training (blue) and testing (orange) losses are high due to randomly initialized parameters. During the first few epochs, the losses drop sharply, showing the model quickly learns patterns while still performing well on unseen data. After around 10 epochs, both losses stabilize near zero with slight oscillations, suggesting

a good balance without overfitting or underfitting. The close alignment of training and testing losses indicates the RNN effectively captures features from the TESS dataset, making it reliable for emotion recognition.

Figure 8 illustrates the accuracy achieved in both training and testing of the RNN model on TESS data for 100 epochs. The x-axis represents the number of epochs, while the y-axis represents the accuracy, with 1 being the highest possible accuracy. To start, the training accuracy (blue line) rises sharply at first, then the testing accuracy (orange line) rises around the same time, demonstrating that the model is capable of quickly grasping the patterns in the data and, quite impressively, is able to generalize. Both accuracy curves reached almost 100% by the 10th epoch and remained around 1.0 throughout the entire training process, as stated. The fact that training accuracy and testing accuracy do not differ much shows that there was less overfitting, indicating the model's ability to learn and generalize well on the TESS dataset for the task of emotion classification.

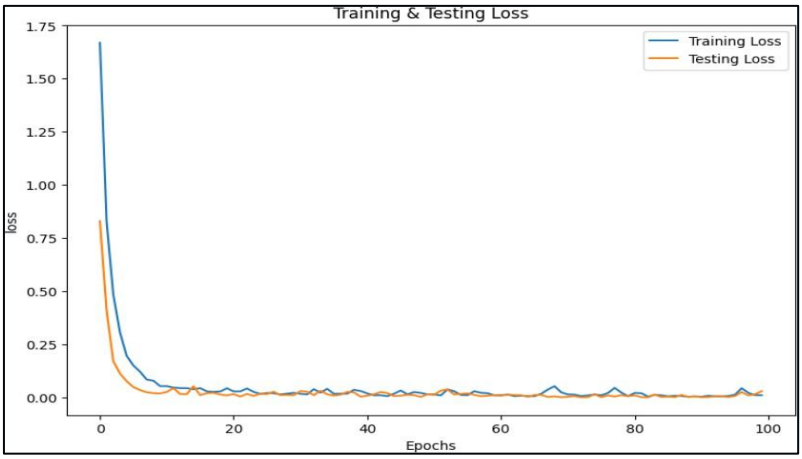

Figure 7. Training and Testing Loss for TESS

Figure 8. Training and Testing Accuracy for TESS

Table 4. Performance Metrics for Different Emotions (TESS)

Emotion	Precision	Recall	F1-score
Angry	1.00	1.00	1.00
Disgust	1.00	0.98	0.99
Fearful	1.00	1.00	1.00
Happy	0.98	1.00	0.99
Neutral	1.00	1.00	1.00
Sad	1.00	1.00	1.00
Surprised	0.98	0.98	0.98
Accuracy			0.99
Macro Avg	0.99	0.99	0.99
Weighted Avg	0.99	0.99	0.99

In Table 4, classification results obtained from running an RNN model on the TESS dataset are depicted, which include F1-score, precision, and recall for every emotion class (Angry, Disgust, Fearful, Happy, Neutral, Sad, and Surprised). Precision shows how accurate the model was in predicting the given emotion, whereas recall indicates how well the model was able to collect all occurrences of the emotion. The F1-score is a measure that takes both precision and recall into account and, therefore, can be a better indicator of performance. while precision, recall, and F1 score are close to 1.00 across categories, reflecting strong classification performance of the model. Overall, the macro and weighted averages for F1 scores of 0.99 illustrate the robustness and generalization of the RNN observed when it is used for emotion recognition on the TESS dataset.

4.4 Test and Train Results of EMO-DB

As portrayed in Figure 9, the graph shows how training and testing loss are reduced over 100 epochs on the Emo-DB data using an RNN. At the start, both losses were high, with training loss around 10 and testing loss close to 2, suggesting a large model fitting error. After 20 epochs, both training loss and testing loss show a gradual decline and reach minimum values close to zero with some fluctuation, which signifies that the model maintained good performance throughout without heavy overfitting. The closeness of the two curves suggests that the RNN generalizes well on unseen samples of the Emo-DB data.

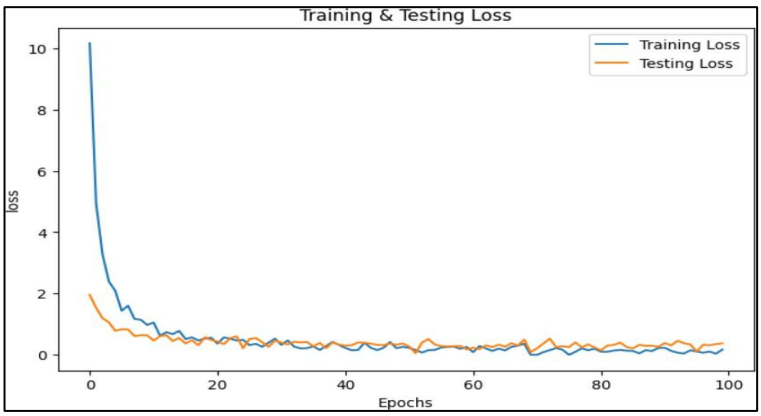

Figure 9. Training and Testing Loss for EMO-DB

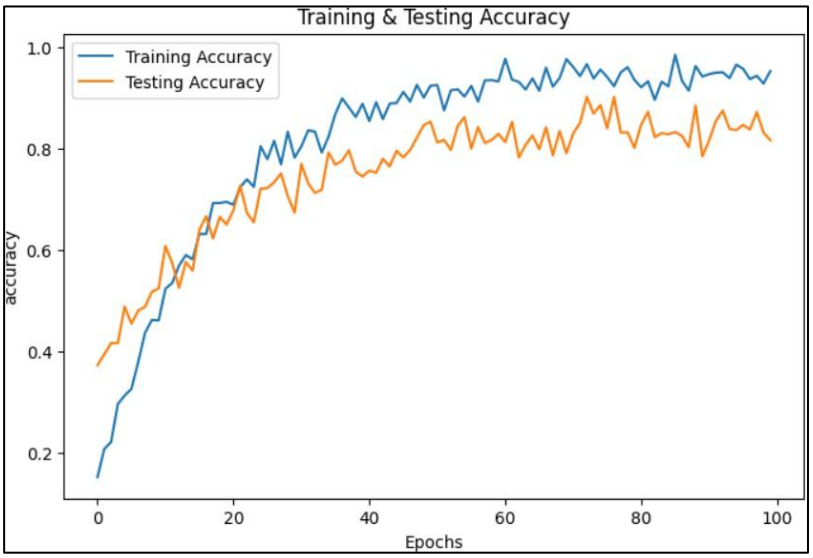

Figure 10. Training and Testing Accuracy for EMO-DB

Figure 10 shows the training and testing accuracy and loss curves corresponding to the performance of the RNN model built and trained for 100 epochs on the EMO-DB database. The training accuracy curve (shown in blue) is observed to exceed 90%, whereas the testing accuracy curve (shown in orange) settles at approximately 70%, signifying that the model learns the available training data well but does not have much capability to predict the outcomes for new data. Similarly, the training loss shows a steep drop and remains at a low value, while the testing loss levels out at a higher value with minor changes.

Table 5. Performance Metrics for Different Emotions (EMO-DB)

Emotion	Precision	Recall	F1-score
Angry	0.82	0.97	0.89
Boredom	0.78	0.74	0.81
Disgust	0.89	0.63	0.79
Fearful	0.72	0.81	0.75
Happy	0.80	0.67	0.74
Neutral	0.79	0.74	0.77
Sad	0.87	1.00	0.93
Accuracy			0.81
Macro Avg	0.81	0.79	0.79
Weighted Avg	0.81	0.80	0.80

Table 5 illustrates a table containing classification results of an RNN model applied to the EMO-DB: Berlin Emotional Speech Database, targeting seven emotions: Fearful, Angry, Neutral, Boredom, Disgust, Sad, and Happy. For each emotion, metrics such as recall, F1-score, and precision are outlined. Model performance is at its best for the “Sad” class, with an F1 score of 0.93; however, lower effectiveness is observed for the “Fearful” and “Happy” classes, with F1 scores of 0.75 and 0.74, respectively. The model's performance measures 0.81 in terms of overall accuracy, with precision measures’ macro and weighted means at 0.81, while F1 measures mean around 0.79 to 0.80, which is a fair performance

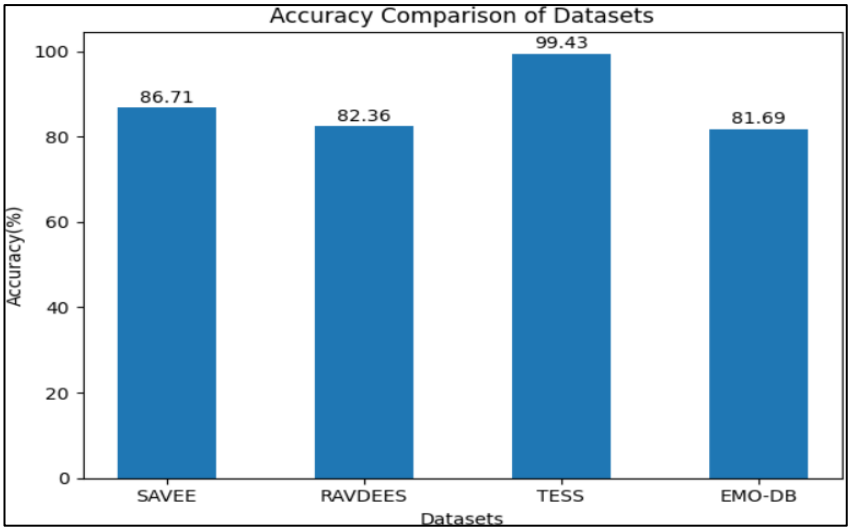

Figure 11. Comparison of Datasets with their Accuracies (%)

The graph in Figure 11 shows the RNN's accuracy across emotional speech datasets SAVEE, RAVDESS, TESS, and EMO-DB. TESS achieved the highest accuracy at 99.43%, indicating that the model performed exceptionally well on it. RAVDESS and SAVEE followed with 82.36% and 86.71%, respectively, while EMO-DB had the lowest at 81.69%, suggesting that it was more challenging for the RNN. These differences highlight how factors like dataset size, number of speakers, and clarity of emotional expression impact performance. Overall, the results emphasize that dataset selection plays a crucial role in emotion recognition tasks.

4.5 Test and Train results of CLAF-SER

As shown in Figure 12, the losses during the training and testing of the CLAF-SER model are plotted over 30 epochs, with the training loss gradually decreasing from a higher initial value (around 1.6) to near-zero, demonstrating successful optimization. The testing loss represents a similar downward trajectory, but stabilizes at a plateau of higher values (around 0.2–0.4), suggesting successful generalization with very little overfitting. The x-axis corresponds to the epochs (0 to 30), while the y-axis shows the loss values, The convergence trend indicates stable adversarial training and strong cross-lingual feature learning. The small gap between the curves indicates a reasonably well-regularized model performance.

As depicted in Figure 13, the training and testing accuracy curves for the CLAF-SER model throughout thirty epochs show increasing values of training accuracy. Such a trend points toward an effective learning process, while testing accuracy displays a pattern of increase, though with minor fluctuations or a tendency to level off; this pattern speaks well of generalization without considerable overfitting. On the x-axis, epochs were charted against the training performances from 0 to 30, while percentage accuracy on the y-axis measures the performance of the model on the training and validation datasets over time. Observations show that both curves remain fairly close to one another; thus, the learning process is balanced and demonstrates a powerful ability to recognize emotions cross-linguistically.

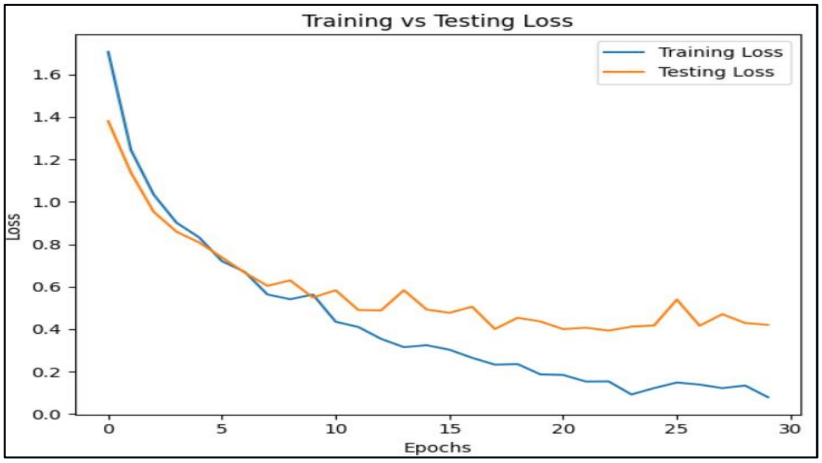

Figure 12. Training and Testing Loss for CLAF-SER

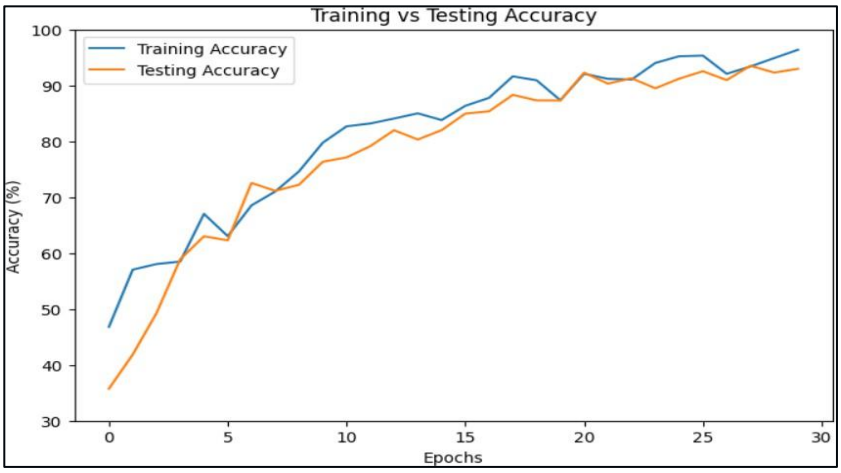

Figure 13. Training and Testing Accuracy for CLAF-SER

The classification performance of CLAF-SER is shown in Table 6 in terms of precision, recall, and F1-score across the different emotions. With an F1-score of 0.93, Calm scores the highest, closely followed by Sad at 0.91, implying that the classification for the former two emotions is better than for others. The lowest F1-score is for Angry (0.80) and Happy (0.81), indicating that these two emotions are not easy to discriminate from other emotions. The average accuracy is excellent at 92% and shows the robustness of this model across multiple datasets for emotion recognition. Furthermore, the macro-average and weighted-average F1-measures were calculated at 0.87 and 0.88, respectively, providing evidence of balanced performance across classes.

Table 6. Performance Metrics for Different Emotions Using the CLAF-SER Model

Emotion	Precision	Recall	F1-score
Neutral	0.92	0.90	0.90
Calm	0.93	0.93	0.92
Happy	0.85	0.81	0.83
Sad	0.94	0.88	0.91
Angry	0.80	0.86	0.85
Fearful	0.88	0.95	0.94
Disgust	0.86	0.85	0.84

Surprised	0.87	0.92	0.89
Accuracy			0.92
Macro Avg	0.88	0.89	0.89
Weighted Avg	0.88	0.88	0.88

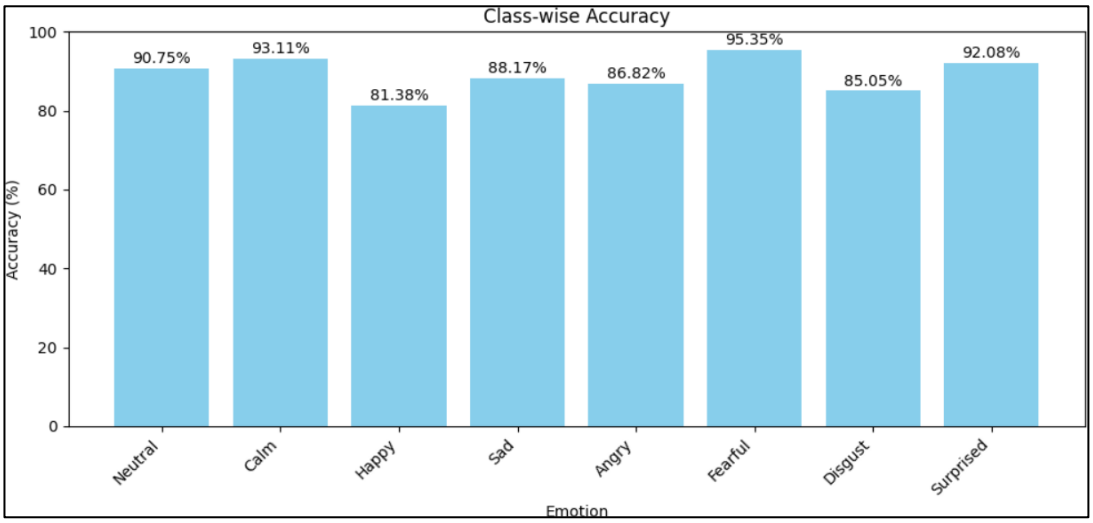

Figure 14. ClassWise Acuracies for CLAF-SER

According to Figure 14, the chart illustrates the class-wise accuracy of CLAF-SER across different emotions. The x-axis covers categories such as Neutral, Calm, Happy, Sad, Angry, Fearful, Disgust, and Surprised, while the y-axis shows accuracy percentages. The model performs best on Fearful (95.35%), followed by Calm (93.11%) and Surprised (92.08%), indicating strong recognition of these emotions. In contrast, Happy records the lowest accuracy at 81.38%, suggesting some difficulty in distinguishing it. Overall, most categories achieve over 90%, highlighting CLAF-SER's effectiveness in cross-lingual emotion recognition.

The RNN-based approach also demonstrates clear advantages over existing methods in terms of both accuracy and training time. For example, while Reference 1 reports accuracies ranging from 57.50% to 99.81%, the proposed RNN achieves comparable or better results with faster training. Similarly, Reference 2's IMEMD-CRNN approach attains high precision with TESS (100%) and Emo-DB (93.54%), but the proposed model balances accuracy with greater efficiency. References 3 and 5 also reach over 90% accuracy using 1D-CNN and 1D-CNN-LSTM-GRU combinations, yet their methods involve significantly higher computational complexity compared to the lightweight RNN used here.

Table 7. Comparison of Methods, Datasets, and Accuracies of Prior Work

Reference	Method	Dataset	Accuracy
[1]	LSTM+Attention+CNN-2D	RAVDESS(R)	74.44%
		SAVEE(S)	57.50%
		TESS(T)	99.81%
		Customized (R+S+T)	90.19%
[2]	IMEMD-CRNN	Emo-db	93.54%
		TESS	100%
[3]	1DCNN	Emo-db	96.6%
		SAVEE	92.6%

		RAVDESS	91.4%
[4]	SVM	Emo-db	85.8%
		IEMOCAP	65.2%
[5]	1D-CNN-LSTM-GRU	TESS	99.46%
		Emo-db	95.42%
		RAVDESS	95.62%
		SAVEE	93.22%
		CREMA-D	90.47%
[6]	SVM	EO (RAVDESS)	89.64%
		CS (RAVDESS)	92.71%
[8]	SVM, GA-CS Algorithm	Emo-db	97.57%
		SAVEE	80%
		PDREC	91.46%
[15]	EmoTech (BiLSTM + CNN)	IEMOCAP	83.52%
[16]	Parallel Model	Wav2Vec 2.0 (IEMOCAP)	72.66%
		HuBERT 2.0 (IEMOCAP)	71.03%
[17]	BRHAMO	RAVDESS	93.8%
		SAVEE	85.4%
		ANAKE	89.8%
[18]	FFREWT-based DNN	EMO-DB	89.35%
		EMOVO	84.69%
		TESS	100.00%
[19]	MFGCN	IEMOCAP	77.3%
		RAVDESS	85.7%
[20]	Multitask Transformer	IEMOCAP, MSP-IMPROV, EMO-DB	(Audio + Text) 46.0%
[21]	AGA-CL	MSP-Podcast(M) (M→B)	58.14%
		BIIC-Podcast(B) (B→M)	55.49%
		Dusha(D) (M→D)	50.91%
[22]	ACO-SVM	EMO-DB	91.5%
		CASIA	88.7%
[23]	DNN	RAVDESS	77.54%
		EMO-DB	87.48%
		SAVEE	79.64%
[24]	Vesper-4	(IEMOCAP, MELD, CREMA-D)	68.4%, 50.1%, 73.4%
	Vesper-12	(IEMOCAP, MELD, CREMA-D)	70.7%, 53.5%, 77.2%
	WavLM Base	(IEMOCAP, MELD, CREMA-D)	65.9%, 49.9%, 59.9%
	WavLM Large	(IEMOCAP, MELD, CREMA-D)	70.6%, 54.2%, 75.7%
Proposed	RNN	RAVDESS(R)	82.36%
		SAVEE(S)	86.71%
		EMO-DB(E)	81.69%

		TESS(T)	99.43%
	CLAF-SER	Custom Data (R+S+E+T)	92.96%

Additionally, if we observe Table 7 carefully, the SVM-based models reported in References 4, 6, and 8 also established a performance trend with high accuracy, achieving up to 97.57% on Emo-DB and 92.71% on CS (RAVDESS). In contrast, the proposed RNN model was successful without much training, achieving an accuracy of TESS on 99.43%. This method of tradeoff between computational power and accuracy makes this approach ideal for scenarios where speed and real-time changes are required. The combination of features with RNN achieves reliability under varying datasets due to complexities, and simultaneously addresses the issues of speed within emotion recognition tasks.

The CLAF-SER model is highly effective for cross-linguistic speech emotion recognition, achieving 92.96% accuracy on a combined dataset of RAVDESS, SAVEE, EMO-DB, and TESS. Unlike single-dataset models such as 1D-CNN-LSTM-GRU (95.62% on RAVDESS) or IMEMD-CRNN (93.54% on Emo-DB), it is designed for multilingual generalization. Using multi-head attention, adversarial domain adaptation, and phonetic feature extraction, CLAF-SER handles language variability better than SVMs (max 92.71%) and deep models like BRHAMO (93.8% on RAVDESS but 85.4% on SAVEE). With BiLSTM for temporal modeling and a Gradient Reversal Layer (GRL) for language-invariant features, it maintains balanced performance across datasets without extensive tuning. This gives it an advantage over models prone to overfitting, like 1D-CNN (96.6% on Emo-DB but $\leq$92.6% elsewhere),and those that don't scale well, such as Vesper-12 (70.7% on IEMOCAP). Even compared to multitask transformers using audio and text (46%), CLAF-SER shows superior adaptability for real-world multilingual tasks. Its interpretable attention weights and adversarial training also add transparency and robustness, setting it apart from typical black-box models while sustaining high accuracy.

4.6 Robustness Under Noisy Conditions

The CLAF-SER model is designed with some degree of noise robustness, thanks to data augmentation during training. The paper introduces a small amount of Gaussian noise (σ = 0.002) to randomly selected audio samples to simulate real-world conditions and variability in speech. While this helps the model learn to generalize across different acoustic scenarios, in this paper haven't yet evaluated its performance under systematically controlled noise levels, such as varying signal-to-noise ratios (SNR). This is an area we plan to explore in future work, possibly by using adversarial noise training or incorporating denoising techniques. Still, the current augmentation strategy already adds useful variation within emotion classes, helping the model become more resilient to common background noise during inference.

5. Conclusion

The CLAF-SER framework demonstrates outstanding performance in cross-lingual speech emotion recognition, achieving 92.96% accuracy on a challenging combined dataset (RAVDESS+SAVEE+EMO-DB+TESS) in just 30 training epochs. Its rapid convergence highlights the model's efficiency compared to classical approaches. By integrating multi-head attention for key feature extraction, adversarial domain adaptation for language-invariant learning, and phonetic encoding for better generalization, CLAF-SER outperforms competitors in both speed and accuracy. Its ability to reach high accuracy with minimal training while

maintaining robustness across multiple datasets makes it ideal for real-time applications where computational efficiency is crucial. CLAF-SER strikes a balance between accuracy, cross-lingual generalization, and low computational cost, making it not only a strong research contribution but also a practical solution for a wide range of speech processing tasks.

CLAF-SER achieves an overall inference latency of around 130–210 milliseconds per utterance in real-time settings. Most of this comes from the feature extraction step using Librosa, while the model itself, consisting of BiLSTM, multi-head attention, and phonetic layers, runs efficiently on a GPU in about 10 milliseconds. Though not designed for streaming, CLAF-SER handles short buffered audio segments (1–2 seconds) smoothly. Further latency reductions are possible by optimizing the feature pipeline or using lightweight, on-device models like Wav2Vec.

To maintain accuracy on edge devices, CLAF-SER is built with an efficient design using lightweight features, BiLSTM, and attention layers. Deployment can be further optimized through post-training quantization, pruning, or knowledge distillation, reducing size and computation with little performance loss. Exporting the model via ONNX or TorchScript also ensures compatibility with mobile and embedded platforms for smooth real-time use. The proposed RNN is not suitable and is not a viable solution to the problem, while in terms of performance, it achieves an accuracy of 99.43%, which is much more suitable for the TESS data set compared to the other data sets. The RNN implemented in this paper was just an experiment to understand how well an RNN could perform in speech emotion recognition.

References

[1] Singh, Jagjeet, Lakshmi Babu Saheer, and Oliver Faust. ”Speech emotion recognition using an attention model.” International Journal of Environmental Research and Public Health, 20(6), 2023: 5140.

[2] Sun C, Li H and Ma L. ”Speech emotion recognition based on improved masking EMD and convolutional recurrent neural network.” Frontiers in Psychology, 13, 2023: 1075624.

[3] Alluhaidan Ala Saleh, Oumaima Saidani, Rashid Jahangir, Muhammad Asif Nauman, and Omnia Saidani Neffati. ”Speech emotion recognition through hybrid features and convolutional neural network.” Applied Sciences, 13(8), 2023: 4750.

[4] Saumard Matthieu ”Enhancing Speech Emotions Recognition Using Multivariate Functional Data Analysis.” Big Data and Cognitive Computing, 7(3), 2023: 146.

[5] Md Rayhan Ahmed, Salekul Islam, AKM Muzahidul Islam, and Swakkhar Shatabda. ”An ensemble 1D-CNN-LSTM-GRU model with data augmentation for speech emotion recognition.” Expert Systems with Applications, 218, 2023: 119633.

[6] Bagadi, Kesava Rao, and Chandra Mohan Reddy Sivappagari. ”An evolutionary optimization method for selecting features for speech emotion recognition.” 21(1), 2023: 159-167.

[7] Wani, Taiba Majid, Teddy Surya Gunawan, Syed Asif Ahmad Qadri, Mira Kartiwi, and Eliathamby Ambikairajah. ”A comprehensive review of speech emotion recognition systems.” IEEE Access, 9, 2021: 47795-47814.

[8] Langari, Shadi, Hossein Marvi, and Morteza Zahedi. ”Efficient speech emotion recognition using modified feature extraction.” Informatics in Medicine Unlocked , 20, 2020: 100424.

[9] Khalil, Ruhul Amin, Edward Jones, Mohammad Inayatullah Babar, Tariqullah Jan, Mohammad Haseeb Zafar, and Thamer Alhussain. ”Speech emotion recognition using deep learning techniques: A review.” IEEE Access, 7, 2019: 117327-117345.

[10] Akc¸ay, Mehmet Berkehan, and Kaya Og˘uz. ”Speech emotion recognition: Emotional models, databases, features, preprocessing methods, supporting modalities, and classifiers.” Speech Communication, 116, 2020: 56-76.

[11] Livingstone SR, Russo FA (2018) TheRyerson Audio-Visual Database of EmotionalSpeech and Song (RAVDESS): A dynamic,multimodal set of facial and vocal expressions in North American English. PLoS ONE 13(5):e0196391.

[12] Burkhardt, F., Paeschke, A., Rolfes, M., Sendlmeier, W.F., Weiss, B. (2005) A database of German emotional speech. Proc. Interspeech 2005, 1517-1520, doi: 10.21437/Interspeech.2005-446

[13] S. Haq and P.J.B. Jackson, ”Multimodal Emotion Recognition”, In W. Wang (ed), Machine Audition: Principles, Algorithms and Systems, IGI Global Press, ISBN 978-1615209194, chapter 17, pp. 398-423, 2010.

[14] Pichora-Fuller, M. Kathleen; Dupuis, Kate, 2020, ”Toronto emotional speech set (TESS)”, https://doi.org/10.5683/SP2/E8H2MF, Borealis, V1

[15] Avro, Shamin Bin Habib, Taieba Taher, and Nursadul Mamun. ”EmoTech: A Multi-modal Speech Emotion Recognition Using Multi-source Low-level Information with Hybrid Recurrent Network.” arXiv preprint arXiv:2501.12674 (2025).

[16] Bautista, John Lorenzo, and Hyun Soon Shin. ”Speech Emotion Recognition Model Based on Joint Modeling of Discrete and Dimensional Emotion Representation.” Applied Sciences 15.2 (2025): 623.

[17] Agrawal, Akshat, and Anurag Jain. ”Brhamo: metaheuristic optimization algorithm for speech emotion recognition using spectral and hybrid features.” Evolutionary Intelligence 18.1 (2025): 4.

[18] Mishra, Siba Prasad, Pankaj Warule, and Suman Deb. ”Fixed frequency range empirical wavelet transform based acoustic and entropy features for speech emotion recognition.” Speech Communication 166 (2025): 103148.

[19] Qi, Xin, et al. ”MFGCN: Multimodal fusion graph convolutional network for speech emotion recognition.” Neurocomputing 611 (2025): 128646.

[20] Ahn, Chung-Soo, et al. ”Multitask Transformer for Cross-Corpus Speech Emotion Recognition.” IEEE Transactions on Affective Computing (2025).

[21] Upadhyay, Shreya G., et al. ”Phonetically-Anchored Domain Adaptation for Cross-Lingual Speech Emotion Recognition.” IEEE Transactions on Affective Computing (2025).

[22] Kang, Xueliang. ”Speech emotion recognition algorithm of intelligent robot based on ACO-SVM.” International Journal of Cognitive Computing in Engineering 6 (2025): 131-142.

[23] [23] Mishra, S.P., Warule, P. & Deb, S. Speech emotion recognition using MFCC-based entropy feature. SIViP 18, 153–161 (2024). https://doi.org/10.1007/s11760-023-02716-7

[24] W. Chen, X. Xing, P. Chen and X. Xu, ”Vesper: A Compact and Effective Pretrained Model for Speech Emotion Recognition” in IEEE Transactions on Affective Computing, vol. 15, no. 03, July-Sept. 2024, 1711-1724. doi: 10.1109/TAFFC.2024.3369726.

[25] Deepika, C., & Kuchibhotla, S. (2023). Design an Optimum Feature Selection Method to Improve the Accuracy of the Speech Recognition System. SN Computer Science,4(5), 655.

[26] Deepika, C., & Kuchibhotla, S. (2024). Deep-CNN based knowledge learning with Beluga Whale optimization using chaogram transformation using intelligent sensors for speech emotion recognition. Measurement: Sensors, 32, 101030.